DOCUMENT RESUME

ED 135 854 TM 006 085

AUTHOR Porter, D. Thomas
TITLE The Development of a Computerized System for the Estimation of Reliability for Measurement Systems Employing Interval or Ratio Data.
PUB DATE [Apr 77]
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Computer Programs; Item Analysis; *Reliability; *Statistical Analysis; *Test Reliability
IDENTIFIERS *PIAS

ABSTRACT

Critical to precise quantitative research is reliability estimation. Researchers have limited tools, however, to assess the reliability of evolving instruments. Consequently, cursory assessment is typical and in-depth evaluation is rare. This paper presents a rationale for and description of PIAS, a computerized instrument analysis system. PIAS makes two major contributions to measurement theory and practice: (1) PIAS is a collection of most of the routines necessary for such analyses in one user-oriented package, and (2) PIAS provides unique output allowing the user to identify the most efficient combination of items; i.e., the smallest number of items with the highest reliability. A utilization agreement and order form for PIAS is included. (Author/RC)

THE DEVELOPMENT OF A COMPUTERIZED SYSTEM FOR THE ESTIMATION OF RELIABILITY FOR MEASUREMENT SYSTEMS EMPLOYING INTERVAL OR RATIO DATA

by

D. Thomas Porter

State University of New York, Buffalo

Critical to precise quantitative research is reliability estimation. Researchers have limited tools, however, to assess the reliability of evolving instruments. Consequently, cursory assessment is typical and in-depth evaluation is rare. This paper presents a rationale for and description of PIAS, a computerized instrument analysis system. PIAS makes two major contributions to measurement theory and practice: (1) PIAS is a collection of most of the routines necessary for such analyses in one user-oriented package, (2) PIAS provides unique output allowing the user to identify the most efficient combination of items; i.e., the smallest number of items with the highest reliability.

Paper presented to the Annual Conference of the American Educational Research Association, New York City, April, 1977.

The Development of a Computerized System for the Estimation of Reliability for Measurement Systems Employing Interval or Ratio Data

by D. Thomas Porter

A cardinal article of faith in measurement theory and practice is reliability estimation. The theoretical and practical value of any research finding is linked inextricably to the internal and external validity of the methodology employed to extract that finding. Validity must in turn be supported by, among other things, a reliable measurement system. Thus when a researcher fails to support the reliability of his measurements, the validity of his conclusions in questionable and their value nondiscernable. In short, the determination of reliability is a fundamental requisite for producing research which has any practical and/or theoretical value. A researcher is left at a disadvantage, however, as there are few, if any, complete, user-oriented, and efficient computer packages for instrument analysis. Consequently, researchers are prone to conduct cursory evaluations of their instruments.

Personal experience of the author suggests that cursory instrument evaluations are often common because extant programs are accessed independently, and if all these programs are employed, their combined output is typically deficient. To conduct a complete evaluation, a researcher must access six or seven different computer packages and programs. Unfortunately, the path of least resistance is usually taken. An average inter-item correlation coefficient is calculated, plugged into Nunnally's formula, and that is it. Such practice is hardly sound measurement technique.

The fundamental premise of PIAS is that the simple computation of a reliability coefficient is an insufficient estimation of an instrument's reliability, nor in any sense is such a complete instrument analysis. Actually several peripheral indices are used to illustrate and/or support the reliability of a measurement system, Correlations with the total score, item discrimination indices, beau coup factor analyses, and split-half reliability checks are common practice. When all of these data are gathered, however, two problems remain. First, several "canned" programs must be accessed, and if they fail to give sufficient information, then additional, time-consuming, and original programs must be written, documented, and validated. Even then, the integration of these outputs is often tedious and inefficient. Second, this information does not answer the following questions:

a) Could higher reliability be achieved with a fewer, select combination of tiems?
b) Could the same reliability be achieved with a fewer, select combination of items?

Central to measurement systems development is measurement efficiency. This goal is particularly important in research where respondent time is at a premium. When new instruments are being tested for the first time, measurement efficiency is even more important. If the same (or higher) reliability and validity can be achieved with 20 items as with 40, then significant research resources can be saved. Normally, a decrease in the number of items causes a decrease in reliability. Such is not always the case, however, as deletions of certain items may, in fact, increase reliability. Unfortunately, current programs and algorithms do not provide such information. With the goals of complete instrument analysis and instrument efficiency in mind, PIAS was written, documented, and validated.

Reliability

Metaphorically, reliability is a measure of the extent to which a set of scores are consistent over time and consistent internally. Theoretically, reliability is the ratio between true scores and error. Accordingly, reliability is a function of two factors: internal consistency and stability. Internal consistency is the extent to which components are consistent with each other. For example, each of the questions' responses on an attitude questionnaire should correlate moderately with each other if the questionnaire is consistent internally. Internal consistency is examined typically by "split-half" correlations, average inter-item correlations, correlations with the total score, or the ability of an individual item to discriminate significantly between a group of high and low scorers (as defined by the total score).

Stability is the extent to which the scores on a measurement system can be produced again at another administration of the system. Stability (commonly called "test-retest" reliability) is usually operationalized by administering the measure at one time and correlating the responses with responses at another time. The adequacy of this procedure is dependent upon several factors. One, the interval (s) of time between administrations is (are) critical. Small intervals allow respondents to remember their former responses and thus artifically inflate the correlation (s) between administrations. Two, stability estimates assume the construct being measured is a trait construct; i.e., corresponding constructs and measures are stable over time. State constructs, on the other hand, and their measures are designed to reflect sensitive environmental changes on purpose. Unless the researcher can control explicitly the environment of the administrations, the stability of state constructs and measures may be difficult to obtain. Three, the design of a study may place more emphasis upon the stability of a measurement system. Any research which employs a measure of change (e.g., from a pre to a posttest) assumes that differences occurring over time are not a function of the instability of the measurement system and are a function of the independent variable.

Stability is used in this context instead of "test-retest" reliability for a very important reason. "Test-retest" procedures may give the researcher a false sense of confidence about a measure's stability. If the measure and its construct are to be generalized to more than two points in time, then stability assessment should comprise more than the typical two administration paradigm. When the measure is used as a predictor of other measures or constructs for a span of several years (e.g., student placement in college), then multiple administrations are absolutely critical.

Overall then, reliability is a function of a measurement system's internal consistency and its stability. Nunnally (1967, page 193) has operationalized this relationship mathematically. He concludes that the formula below "cannot be overemphasized in its importance for measurement theory."

$$\text{Reliability} = \frac{Kr}{1 + (K - 1)r}$$

In the above theorem/formula, Nunnally has incorporated both components of reliability. The K represents the number of items in the measurement system; whereas, r the average inter-item correlation coefficient between components. Stability is reflected in K and internal consistency is reflected in r. The larger the K value, the smaller the extent any one item's aberrations can change the overall scores at a later administration and thus, the higher the stability of the instrument. The higher the value of r, the greater the degree of internal consistency between items. This theorem assumes that the measurement system is designed to measure one construct. If it measures more than one construct, then the reliability will be lower as the degree of independence between constructs grows. Accordingly, sub-constructs or sub-tests are assessed individually as to their reliability.

The reliability coefficient which results will range in value from 0.0 to 1.0 with 0.0 indicating no reliability and 1.0 indicating perfect reliability. The size of the number of observations taken to estimate reliability does not directly affect the size of the reliability coefficient. If the number of observations is small or selected non-randomly, however, the variability of the items' scores may be small and deflate the coefficient accordingly. In addition, the degree of confidence that the obtained coefficient represents the true reliability is directly proportional to the number of observations collected; the larger the "N" the better.

The "adequacy" of a reliability coefficient is a thorny matter for researchers; for the most part, it is largely subjective. In some cases, the "adequacy" issue is purely academic; a given reliability may be all that is feasible. Whenever possible,

however, the researcher should try to increase reliability. There are some objective criteria which should be considered when asking is my reliability "high enough?" For example, a coefficient of .70 indicates that the instrument accounts for about 49% (the coefficient squared times 100) of the variance within the measurement system; or, in other words, over half of the internal variance is extraneous. In this instance a researcher would normally want to improve the instrument's reliability. One should also realize that low reliability tends to lower the probability that a null hypothesis will be rejected. In other words, low reliabilities have a tendency to conservatize hypothesis testing and reduce the efficiency of the ratio between research resources expended and meaningful results obtained. When prediction of other variables is the purpose of a measurement system, then low reliabilities are even more critical. In this last case inaccurate and imprecise predictions will be more probable.

Of course, reliability is only one part of the total research process. Once sufficient reliability estimates are obtained, the more important questions of validity enter. But without sufficient reliability, validity is, by definition, a moot question. In short reliability and instrument analysis is only a first, but necessary step in scientific research.

Program PIAS

PIAS is a multi-phasic, computerized instrument analysis system. It is user-oriented and requires that the user know how to format the data (tell PIAS where to find the data upon input). It provides five different analyses, each of which give information as to an instrument's reliability. In addition, PIAS provides many descriptive statistics and, when appropriate, a test for additivity of items. A typical use would be the analysis of a Likert-type attitude scale where the user needs to know what items are decreasing reliability, if any, and what is the most efficient cluster of items; i.e., what combination of items gives the highest reliability with the smallest number of items. For a complete list of inputs and outputs, see Tables One and Two.

Phase I of PIAS is descriptive in nature. For each component in the measurement system Phase I provides means, standard deviations, standard errors of the mean, kurtosis and skewness values, probabilities of kurtosis and skewness values, highs, lows, and ranges of each item. If the user desires, distribution plots of each item will also be printed. In addition it provides a correlation matrix, t and p values matrix, and a correlation matrix assessment. In this phase PIAS notes and stores all items which failed to correlate significantly with greater than 60 per cent of the other items. (NOTE: at this point and all others where tests of significance are conducted, adjustments are made

to account for multiple tests. This is accomplished by dividing the user's input alpha by the number of tests to be conducted, Bonferonni).

Phase II is concerned with additivity of items in the measure. Since many scales are comprised of items which are assumed to be equal in importance and are later added together to form a total score, this phase is very important for instrument development. Phase II conducts a test of significance between all meaningful pairs of item means, reports the t and p values, and then a summary table describing which items differed significantly with which other items. Finally, a one-way ANOVA is conducted across the total group of items to give an indication of additivity. Homogeneity of variance indices are provided with this test.

Phase III of PIAS is concerned with what the most efficient combination of items is with respect to reliability. Each item is rank-ordered by its degree of contribution to the reliability coefficient. This calculation is accomplished by a test of significance between the average correlation of one item with all other items and the average inter-correlation of all items. Any item whose correlations with other items is significantly lower than the average inter-item correlation coefficient is so noted with an asterik. Phase III ends with a description of the most efficient combination of items.

Phase IV is concerned primarily with internal consistency. Two forms of output are provided. The first output is correlations with the total score. This analysis is based upon the assumption that the best estimate of the true score for a given case is the total score of all items. Any item which fails to correlate with the total score should be viewed as a questionable item (assuming a uni-dimensional measure). The second output is an indication of each item's ability to discriminate between high scorers and low scorers. Any item which could not do so, would also be questionable from an internal consistency point of view.

Phase V serves a summary function. At this point PIAS calculates what the reliability would be if items failing to meet the criteria specified in Phases II-IV were deleted. Phase V also gives the overall reliability and any additional descriptive statistics on total scores, if appropriate.

PIAS is written in FORTRAN for the Cyber 173 NOS 1.1 operating system at the State University of New York at Buffalo (RUNW compiler). It employs 19 subroutines and requires a field length of 44300 octal plus sufficient field length (core memory) for the FORTRAN compiler. Although developed at the CDC Cyber 173 which has a 60 bit word capacity, PIAS was written to accomodate machines with 32 bit or larger word capacities. The source deck is approximately 2000 cards in length.

Summary and Personal Note

Because reliability estimation is a paramount concern for researchers interested in conducting quality research and because current programs are incomplete and disorganized, PIAS was developed. PIAS allows the user not only to estimate reliability and completely assess his instrument, but also to identify the precise combination of items which give the most efficient reliability. Researchers, like most human beings, often follow paths of least resistance. If editors do not insist upon reliability information, researchers will rarely provide it. Correspondingly, if improving reliability and conducting instrument analyses means several trips to the computer center because of several disjointed programs and analyses, then cursory instrument evaluations will continue to be the norm. The author sincerely hopes users find PIAS useful.

Table One

Inputs for Program PIAS

1) Six character job name for analysis.
2) The number of cases (subjects, n); needed only if input is from cards. n must be greater than 1 and less than 3001.
3) The number of items (components, i) in the measurement system. i must be greater than 1 and less than 101.
4) The alpha the user wants to maintain in the analyses. Alpha must be greater than zero and less than .26.
5) Whether punched card output is desired. If so, whether sums of standardized or non-standardized items are desired.
6) Whether a listing of input data is desired.
7) Whether a heading for the analyses is to be read in and printed.
8) Whether non-standardized items can be logically summed or whether items should be standardized (converted to Z-scores) before summing across items. (If neither, some of the output listed in Table Two will not be produced.)
9) Whether cases in high and low categories of scorers are to be printed.
10) Whether distributions of all items are to be plotted.
11) The source of input (cards, file, or magnetic tape).
12) Effect size for difference testing.
13) Effect size for correlational testing.
14) Discrimination analysis classification factor.
15) A FORTRAN-type format for the data input source.
16) A heading card (optional).

Table Two

Output Products of Program PIAS

Descriptive Data for Each Item of the Instrument:

- A1) Means and standard deviations.
- A2) Sums and standard errors of the mean.
- A3) Skewness and kurtosis values.
- A4) Probabilities of skewness and kurtosis values.
- A5) High, low, and range characteristics.
- A6) Means $\pm$ standard deviations.
- A7) Correlation matrix; t-values and p values associated with same; alpha is adjusted for multiple tests.
- A8) Distribution plots for each item (optional).

Reliability Information for the Instrument:

- B1) Rank order of each item's contribution to the overall reliability coefficient; most efficient group of items.
- B2) Correlation matrix assessment which identifies all items not correlating with any one particular item.
- B3) Assessment of items which differ significantly with each other (optional).
- B4) Analysis of variance for item additivity (optional).
- B5) Correlations of each item with the total score (optional).
- B6) Discrimination analyses for each item (i.e., items' ability to discriminate significantly between groups of high and low scorers, optional).

Peripheral Output:

- C1) Parameter check to ensure all input parameters are legal.
- C2) Heading for the analysis (optional).
- C3) Listing of input data (optional).
- C4) Punched card output of standardized or non-standardized items' sums, identification data, and case numbers (opt.).
- C5) Listing of cases, their summed scores, and identification data in the high and low categories of total scorers. (opt.).
- C6) Reliability coefficients with increased and decreased numbers of items/components.
- C7) Summary of analysis:
  - C7a) Reliability of measurement system, average inter-item correlation coefficient among items.
  - C7b) Reliabilities if items designated by criteria above (B1 to B6) were deleted.
  - C7c) Mean, standard deviation, standard error, kurtosis, skewness, percentiles, quartiles, and distribution plot of the total scores (conditional).

UTILIZATION AGREEMENT AND ORDER FORM FOR PROGRAM PIAS

Program PIAS is intended for use only by non-profit and non-military institutions and individuals. Any deviation from this principle without the expressed written consent of the author is strictly prohibited by the copyright for PIAS and written agreement herein. If use of same constitutes a non-profit use, permission is necessary for legal use of PIAS. For further information write or call:

Dr. D. Thomas Porter
4226 Ridge Lea Road
State University of New York, Buffalo
Buffalo, New York 14226
716-838-3208 or 716-831-1607

Please use the form below for ordering manuals, source decks, or requesting services. Please print or type.

Name_____

Institution_____

Address_____

_____Zip_____

I agree that the use of program PIAS will be for non-profit and non-military research and/or instructional purposes only.

Signed_____

Check desired items:

	QUANTITY	PRICE	TOTAL	
_____ Manual (s)	_____	$4.00@	_____	
_____ Source Deck	1	$60.00	_____	029 or 026 punch (circle one)
_____ Installation*	na	$275.00	_____	

Grand Total $_____

_____ Check or money order enclosed (no cash)

_____ Please bill me.

*Price of installation covers cost of source deck and 5 manuals, but it does not cover travel and per diem expenses which are to be borne by the purchasing institution or individual.